Sidman Goes to College: A Meta-Analysis of Equivalence-Based Instruction in Higher Education

Julia Brodsky¹ · Daniel M. Fienup²

Published online: 26 April 2018

© Association for Behavior Analysis International 2018

Abstract Equivalence-based instruction (EBI) is a pedagogy based on the principles of stimulus equivalence for teaching academically relevant concepts. This self-paced and mastery-based methodology meets many of the instructional design standards suggested by Skinner (1968), adds generative learning by programming for derived stimulus–stimulus relations, and can be particularly useful in the context of a college course in which students must learn numerous concepts. In this article, we provide the first meta-analysis of EBI in higher education. The authors conducted a systematic literature search that identified 31 applied, college-level EBI experiments across 28 articles. Separate meta-analyses were conducted for single-subject and group design studies. Results showed that EBI is more effective than no instruction and an active control and that studies comparing EBI variants show differences between training variants. Future research should increase internal, external, and statistical conclusion validity to promote the mainstream use of EBI in classrooms.

Keywords Equivalence-based instruction · Higher education · Meta-analysis · Relational frame theory · Stimulus equivalence · Systematic review

In 1971, Murray Sidman described the emergence of novel stimulus–stimulus relations, or “equivalences,” when he observed an individual with an intellectual disability demonstrate comprehension (selecting print words in the presence of corresponding pictures) despite never having received direct intervention on this skill. The participant had a preexisting repertoire of matching spoken words (A) with pictures (B) and was taught to match spoken words (A) to written words (C). Through an association with

✉ Daniel M. Fienup
Fienup@tc.columbia.edu

¹ Department of Psychology, Queens College and the Graduate Center, CUNY, New York, NY, USA

² Department of Health and Behavior Studies, Teachers College, Columbia University, 525 W. 120th St., Box 223, New York, NY 10027, USA

spoken words ($A \rightarrow B$, $A \rightarrow C$), the participant was then able to associate pictures and written words ($B \rightarrow C$, $C \rightarrow B$) without additional formal instruction. Due to the physical dissimilarity between the stimuli, stimulus generalization did not explain the emergence of picture and written word stimulus–stimulus relations. Sidman later applied mathematical set theory to describe the emergence of novel stimulus–stimulus relations where each respective stimulus had been paired with a mutual stimulus (stimulus equivalence; Sidman, 1994). A large number of basic and applied studies followed that elucidated the behavioral principles governing equivalence class formation and how these principles can be applied to promote socially relevant skill acquisition (Rehfeldt, 2011).

Researchers have applied these same principles to the design of college-level curricula. College degrees have become increasingly important due to the realized lifetime economic benefits and a college degree becoming a prerequisite for entry-level employment. The increased need for a bachelor's degree suggests that it is important to help college students complete their degrees within the standard 4-year period. A longitudinal study conducted by the US Department of Labor found that 38% of students had some college education but had not completed their bachelor's degree by the time they were 27 years of age (Bureau of Labor Statistics, 2016). Stimulus equivalence applications to higher education address efficiency of instruction challenges that may underpin low student performance scores through a course of study.

Most college classes are taught in large lecture formats (Mulryan-Kyne, 2010), and it has been suggested that these formats rely on aversive control of student performance (Michael, 1991). According to Skinner (1968), effective instruction, broadly speaking, should be individualized, self-paced, allow opportunities for frequent responding, provide frequent feedback, and progress as a learner demonstrates mastery of each lesson. Additionally, instruction should be generative to facilitate emergent learning—saving instructional time and thereby demonstrating efficiency (Critchfield & Twyman, 2014; Keller, 1968). Educational stimulus equivalence applications address basic and generative aspects of instruction.

Equivalence-based instruction (EBI) incorporates the principles of stimulus equivalence¹ (Rehfeldt, 2011; Sidman, 1994) within instructional design to teach academically relevant concepts (Fienup, Hamelin, Reyes-Giordano, & Falcomata, 2011). Typically using match-to-sample (MTS) procedures, EBI teaches learners to treat physically disparate stimuli as functionally interchangeable by training overlapping conditional discriminations. Instructors arrange contingencies to teach the respective conditional discriminations in order and to mastery. EBI is economical and generative (Critchfield & Fienup, 2008; Fields, Verhave, & Fath, 1984; Sidman & Tailby, 1982). By training only two overlapping baseline relations to mastery, an instructor typically observes the emergence of additional derived relations:

¹ As other articles in the present issue make clear, there are other kinds of stimulus relations, and these, like equivalence relations, can provide a foundation for designing instruction. However, because most classroom applications to date have focused on equivalence relations, this term appears to be in common use, and we will stick with convention and use it herein.

Symmetry, transitivity, and equivalence (for more details, see Sidman (1994). A learner who has mastered all baseline and derived relations is said to have formed an equivalence class (Green & Saunders, 1998). Instructors can use EBI to increase the effectiveness—and potentially the efficiency—of learning in higher education. Researchers have observed large academic gains in little time, maximizing training benefits (Fienup, Mylan, Brodsky, & Pytte, 2016). Furthermore, the training benefits, relative to the cost of time engaged in direct instruction, may increase with repeated EBI tutorials. This outcome has been demonstrated by a few studies suggesting participants require less time to complete training with successive EBI tutorials (e.g., Fienup et al., 2016).

Research on the positive educational outcomes produced by college-level EBI applications has been building in the last few decades. Researchers have applied EBI to a variety of academic topics, such as statistics (Albright, Reeve, Reeve, & Kisamore, 2015; Fields et al., 2009), keyboard playing (Hayes, Thompson, & Hayes, 1989) hypothesis testing (Critchfield & Fienup, 2010; Critchfield & Fienup, 2013; Fienup & Critchfield, 2010), algebra and trigonometric functions (Ninness et al., 2006), disability categorization (Alter & Borrero, 2015; Walker, Rehfeldt, & Ninness, 2010), neuroanatomy (Fienup, Covey, & Critchfield, 2010; Reyes-Giordano & Fienup, 2015), and behavior science topics such as single-subject research design (Lovett, Rehfeldt, Garcia, & Dunning, 2011) and the interpretation of operant functions (Albright, Schnell, Reeve, & Sidener, 2016). Although most EBI applications teach concepts using computerized, programmed instruction in laboratory settings, some have also incorporated lectures (Critchfield, 2014; Fienup et al., 2016; Pytte & Fienup, 2012), paper worksheets (Walker et al., 2010), and distance learning platforms (Critchfield, 2014; Walker & Rehfeldt, 2012).

Two previous reviews (Fienup et al., 2011; Rehfeldt, 2011) suggested that EBI is an effective instructional intervention for teaching various skills across a variety of formats to adult learners, although both of these surveys are now dated. Both also used qualitative review methods, and a major purpose of the present article is to derive insights from the quantitative methods of meta-analysis. To help researchers determine the magnitude of a treatment effect on the population, a meta-analysis aggregates effect sizes from a number of studies examining the treatment effects on various samples. An effect size is a standardized metric by which researchers can ascertain the magnitude of treatment outcomes and compare treatment effects across a variety of studies and measures (Field, 2009). Statisticians have suggested that the reporting of effect sizes and their confidence intervals (CIs) may help circumvent some of the limitations of null hypothesis significance testing, confirming the existence of an effect (Cohen, 1994). In the current study, we limited our analysis to studies focusing on college instruction because a sizeable body of research involving this population is now available. For this population, our review sought to answer three primary questions:

1. Is EBI effective?
2. Are there variations of EBI that produce better academic outcomes?
3. Is EBI more effective than alternative instructional strategies?

Quantifying the answers to these three questions can help direct the goals of future EBI research and increase its use in the classroom and in other naturalistic educational settings.

Method

Inclusion and Exclusion Criteria

To be included in our review, studies had to:

- 1) Be written in English;
- 2) Have been published in a peer-reviewed journal;
- 3) Use stimulus equivalence methodology (i.e., train overlapping conditional discriminations, test for derived relations);
- 4) Describe an experiment in which the implementation of EBI was at least one factor of the independent variable;
- 5) Include participants who were college undergraduate or graduate students; and
- 6) Include only stimuli that were academically relevant to college students (this criterion excluded studies, mostly basic studies, in which at least some of the stimuli were arbitrary).

The inclusion criteria did not discriminate between experiments described from the perspectives of stimulus equivalence (Sidman, 1994) and relational frame theory (Hayes, Barnes-Holmes, & Roche, 2001), as both perspectives result in a pedagogy captured by the third inclusion criterion.

Procedure

Article Search The literature review was conducted in three stages between March 2016 and August 2016. Figure 1 summarizes the search stages described in the following sections. If the title of an article met one or more exclusion criteria, we excluded the article without further analysis. If an article's title did not meet any of the exclusion criteria, we then examined the abstract and article text to determine whether the article met the inclusion criteria.

Stage 1: keyword search. We identified three search terms that generated EBI studies fitting the aforementioned criteria: *stimulus equivalence AND college*, *equivalence-based instruction AND college*, and *derived relations AND college*. The first author entered these search terms into both PsycINFO and ERIC ProQuest. She then recorded the number of hits for the three search terms in both databases. Figure 1 shows the number of search hits and included articles for each search term. Subsequently, an independent observer repeated this procedure. Observers agreed on all cases, yielding 100% interobserver agreement (IOA).

Stage 2: article search. The first author then applied the inclusion and exclusion criteria to each of the stage 1 articles. An independent observer applied the inclusion and exclusion criteria to 33% of these articles. Of the 65 unique articles, 13 articles met the criteria for our review and meta-analysis. Observers agreed on all of the 33% of cases analyzed, yielding 100% IOA.

Stage 3: citation and reference search. Stage 3 involved a citation and reference search of the 13 articles found in stage 2. For each of the 13 identified articles, the

Stage 1

Keyword Search		
Keyword	Search Engine	Number of Hits
"Equivalence-Based Instruction" AND "College"	Eric ProQuest	8
	PsychINFO	0
"Derived Relations" AND "College"	Eric ProQuest	10
	PsychINFO	1
"Stimulus Equivalence" AND "College"	Eric ProQuest	24
	PsychINFO	35
Total:		78

Article Search

65 unique articles

13 included

Stage 2

Stage 3

Fig. 1 Method for article search and collection

first author reviewed all articles found in the References sections ($N=210$) and articles that cited the identified article ($N=127$), according to a Google Scholar search. For each newly identified article, the first author applied the inclusion and exclusion criteria, yielding 4 novel articles from the reference search and 10 articles from the citation search. The first author conducted new citation and reference searches for the 14 newly discovered articles that revealed no novel reference search articles and one novel citation search article. This process was repeated again with the one newly discovered article, and no additional novel articles were identified; therefore, we concluded stage 3. IOA was evaluated for 33% of the stage 3 articles, and observers agreed in 96% of cases.

Data Collection In total, the search yielded 28 unique articles containing 31 experiments (see Table 1) that met our inclusion criteria. We coded the 28 articles for the dependent variables listed in Table 2 and coded the 31 experiments for the variables listed in Tables 3 and 4. Data collection IOA between independent observers was determined for 11 experiments (35%) on the variables listed in Tables 3 and 4. To calculate IOA on a per-experiment basis, we divided the number of agreements by the total number of ratings and multiplied that number by 100. The IOA for data collection was 92% (range of 78% to 96%).

Effect Size Calculations The studies included both group comparisons and single-subject designs. Effect sizes were computed for group design studies using Hedges's g (Lipsey & Wilson, 2001), and effect sizes for single-subject designs were calculated using the improvement rate difference (IRD; Kratochwill et al., 2010). Effect size interpretations vary between measures, and therefore, statisticians have categorized them to reflect different relative magnitudes (small, medium, and large) of treatment based on standard recommendations for different types of calculations (Field, 2009; Lipsey & Wilson, 2001). This categorization permits the comparison of effect sizes calculated using different methods. Hedges's g was chosen for group designs because it corrects for unequal sample sizes (Ellis, 2010). Values for Hedges's g were interpreted using the following guidelines: Values equal to or less than 0.5 were considered small, values greater than 0.5 and less than 0.8 were considered medium, and values equal to or greater than 0.8 were considered large (Field, 2009; Lipsey & Wilson, 2001). IRD was chosen because of its sensitivity over other single-subject effect size calculations and its ability to calculate CIs (Parker, Vannest, & Brown, 2009). An IRD value less than 0.50 indicated a small effect, a value between 0.50 and 0.70 indicated a moderate effect, and a value greater than 0.70 indicated a large effect (Parker et al., 2009).

Group Designs Of the 25 experiments that used group designs, five of the corresponding manuscripts provided a complete data set for calculating Hedges's g . We were unable to include McGinty et al. (2012)'s or Walker et al. (2010, experiment 2) in the meta-analysis due to small sample size, which caused a computing error in the meta-analysis software. For the remaining 20 studies, we contacted the respective researchers and asked them to provide the necessary data. In this way, data were obtained for 13 additional experiments. For one additional study (Fields et al., 2009), we were able to extract the relevant data from Fig. 5 of the source article.²

We obtained data for the calculation of Hedges's g from an article's published figures and tables or used researcher-provided raw data to calculate descriptive and inferential statistics. All obtained data were entered into Comprehensive Meta-Analysis (2014), which calculated Hedges's g , 95% CIs and fixed or random effects. The fixed-effect model assumes that the effect measured in the studies analyzed is true. The program calculated fixed effects for primary measures. Primary measures were assumed to represent true effects because they were the direct product of the EBI intervention (Field, 2009). Primary measures were defined as data collected from posttests conducted in the same topography

² Specifically, we copied the published graph and pasted it into Microsoft Excel so that a grid could be superimposed on it. The grid was used to determine values for raw data points. Value-by-value IOA was collected to ensure accurate data extraction, with agreement obtained on 41 of 42 values (98%). For the one disagreement, we reviewed the data point and came to a consensus about the value displayed in the graph.

Table 1 Basic information for included articles and experiments

Reference	Year	Number	Content	Design	Included in meta-analysis	EBI vs. NIC	EBI vs. EBI	EBI vs. AC
Albright et al.	2015	10	Statistics	Group	Yes	Yes		
Albright et al.	2016	10	Operand functions	Group	Yes	Yes		
Alter and Borrero	2015	17	Disorders and disabilities	Group	Yes	Yes		
Critchfield	2014	60	Hypothesis testing	Group	Yes	Yes		
Critchfield and Fienup	2010	27	Hypothesis testing	Group	No	Yes		
Critchfield and Fienup	2013	5	Hypothesis testing	Group	No	Yes		Yes
Fields et al.	2009	21	Statistics	Group	Yes	Yes		
Fienup and Critchfield	2010	10	Hypothesis testing	Group	Yes	Yes		
Fienup and Critchfield	2011	42	Hypothesis testing	Group	Yes	Yes		Yes
Fienup, Covey, and Critchfield	2010	4	Neuroanatomy	SSD	Yes	Yes		
Fienup, Critchfield, and Covey, experiment 1	2009	12	Statistics	Group	Yes	Yes		
Fienup, Mylan, Brodsky, and Pytte	2016	27	Neuroanatomy	Group	Yes	Yes	Yes	
Fienup, Wright, and Fields, experiment 1	2015	43	Neuroanatomy	Group	Yes	Yes	Yes	
Fienup, Wright, and Fields, experiment 2	2015	24	Neuroanatomy	Group	Yes	Yes	Yes	
Hausman et al.	2014	9	Portion size estimation	SSD	No	Yes		
Hayes, Thompson, and Hayes, experiment 1	1989	9	Music	Group	No	Yes		
Hayes, Thompson, and Hayes, experiment 2	1989	9	Music	Group	No	Yes		
Lovett et al.	2011	24	Research design	Group	Yes	Yes		Yes
McGinty et al.	2012	3	Mathematics	Group	No	Yes		
Ninness et al.	2006	8	Mathematics	Group	Yes	Yes		
Ninness et al., experiment 2	2009	4	Mathematics	SSD	Yes	Yes		
O'Neill et al.	2015	26	Skinner's verbal operants	Group	Yes	Yes		Yes

Table 1 (continued)

Reference	Year	Number	Content	Design	Included in meta-analysis	EBI vs. NIC	EBI vs. EBI	EBI vs. AC
Pytte and Fienup	2012	93	Neuroanatomy	Group	No	Yes		
Reyes-Gordano and Fienup	2015	14	Neuroanatomy	SSD	Yes	Yes		
Sandoz and Hebert	2016	24	Statistics	Group	Yes	Yes		
Sella, Ribeiro, and White	2014	4	Research design	SSD	Yes	Yes		
Trucil et al.	2015	3	Portion size estimation	SSD	Yes	Yes		
Walker and Rehfeldt	2012	11	Research design	Group	Yes	Yes		
Walker, Rehfeldt, and Ninness, experiment 1	2010	13	Disorders and disabilities	Group	Yes	Yes		
Walker, Rehfeldt, and Ninness, experiment 2	2010	4	Disorders and disabilities	Group	Yes	Yes		
Zinn, Newland, and Ritchie	2015	61	Drug names	Group	No			Yes

EBI vs. *NIC* indicates that a study compared EBI scores to no-instruction control scores, *EBI* vs. *EBI* indicates that a study compared scores from variations of EBI, and *EBI* vs. *AC* indicates that a study compared EBI scores to active instructional control scores

EBI = equivalence-based instruction, *NIC* = no-instruction control, *AC* = active control, *SSD* = single-subject design

Table 2 Background information for included articles

Journals	Number of articles (%)
<i>The Analysis of Verbal Behavior</i>	1 (4)
<i>European Journal of Behavior Analysis</i>	1 (4)
<i>Journal of Applied Behavior Analysis</i>	16 (57)
<i>Journal of Behavioral Education</i>	2 (7)
<i>Journal of the Experimental Analysis of Behavior</i>	1 (4)
<i>The Experimental Analysis of Human Behavior Bulletin</i>	2 (7)
<i>The Journal of Undergraduate Neuroscience Education</i>	1 (4)
<i>The Psychological Record</i>	4 (14)
Framework	Number of articles (%)
Relational frame theory	4 (14)
Stimulus equivalence	24 (86)

as training. The random-effect model allows for the measured effect to differ between studies (Field, 2009). The program calculated random effects for secondary measures because behavior on these measures was allowed to vary more than behavior measured in the trained topography across studies. Secondary measures were defined as those measuring generalization (across time, with novel stimuli, or across a novel response topography).

We generated forest plots by taking the values of Hedges's g and confidence intervals from Comprehensive Meta-Analysis (2014) and plotting the data.

Single-Subject Designs Using published graphs or raw data from the six single-subject designs, we calculated IRD (Kratochwill et al., 2010) for five of the six single-subject design experiments (see Table 1). We omitted Hausman, Borrero, Fisher, and Kahng's (2014) experiment because the data, which focused on reducing variability, were not amenable to IRD calculations.

The first step in calculating IRD is to determine the number of data points that overlap in baseline and training or intervention. This process is done separately for baseline and training. To calculate IRD, we subtracted the percentage of baseline data that overlapped with training data points from the percentage of training data points that overlapped with baseline data points (Parker et al., 2009). IRD effect sizes were determined using an online CI calculator (VassarStats; http://www.vassarstats.net/prop2_ind.html) and are reported on a scale from 0 to 1.00.

To calculate omnibus IRD, we added together values in each of the following four separate groups, per category of study: (a) training data points that did not overlap with baseline, (b) total number of data points in training, (c) baseline data points that did overlap with training, and (d) total number of data points in baseline. We then divided the training data points that did not overlap with baseline by the total number of data points in training and divided the baseline data points that did overlap with training by the total number of data points in baseline. Finally, we subtracted the baseline quotient from the training quotient to determine omnibus IRD. VassarStats was used to find CIs for IRD values. We calculated the following omnibus IRD values: per experiment, for all primary measures, and for all secondary measures.

Table 3 Method information for included experiments

Participant characteristics	Experiments reported			
	Yes		No	
	No.	%	No.	%
Age	14	45	17	55
Race	4	13	27	87
Gender	12	39	19	61
SES	0	0	31	100
SAT score	0	0	31	100
ACT score	3	10	28	90
GPA	7	23	24	77
Participant level of schooling	Number of experiments (%)			
Not reported	4 (13)			
Graduate only	5 (16)			
Undergraduate and graduate	2 (6)			
Undergraduate only	20 (65)			
Participant compensation				
Extra credit	8 (26)			
Money	5 (16)			
Extra credit and money	4 (13)			
Course requirement or course credit	12 (39)			
Other	1 (3)			
None	1 (3)			
Setting				
Classroom	5 (16)			
Distance education (e.g., Blackboard, Moodle)	4 (13)			
Laboratory	20 (65)			
Not mentioned	2 (6)			
Experimental design				
Group	25 (81)			
Between	3 (12)			
Within	12 (48)			
Mixed	10 (40)			
Single subject	6 (19)			
Between	0 (0)			
Within	6 (100)			
Mixed	0 (0)			
Campbell and Stanley (1963) classification				
Time series experiment (quasi-experimental design)	6 (19)			
Pretest–posttest control group design (true experimental)	12 (39)			

Table 3 (continued)

Participant characteristics	Experiments reported			
	Yes		No	
	No.	%	No.	%
One-group pretest–posttest design (preexperimental)	12	(39)		
Posttest-only control group design (true experimental)	1	(3)		
Training structure				
One to many (OTM)	13	(42)		
Many to one (MTO)	0	(0)		
Linear series (LS)	7	(23)		
Mixed	11	(35)		
Training protocol				
Simultaneous	12	(39)		
Simple to complex	13	(42)		
Other or mixed	6	(19)		
Testing format				
Written topographical	6	(19)		
Written MTS	13	(42)		
Computer-based topographical	3	(10)		
Computer-based MTS	21	(68)		
Spoken topographical	5	(16)		
Portion size estimation	2	(6)		
Other	2	(6)		

SES = socioeconomic status, *SAT* = Scholastic Aptitude Test, *ACT* = American College Testing, *GPA* = grade point average, *MTS* = match to sample, *OTM* = one to many, *MTO* = many to one, *LS* = linear series

Table 4 Data reported for included experiments

Measure	Experiments reported			
	Yes		No	
	No.	%	No.	%
Interobserver agreement (IOA)	15	48	16	52
Treatment integrity (TI)	4	13	26	84
Percentage correct on derived relations probes	27	87	4	13
Yield	3	10	28	90
Time	10	32	21	68
Trials or blocks to criterion	27	87	4	13
Social validity	9	29	22	71

Results

Description of EBI Experiments

Table 1 lists basic information about the 28 articles and 31 experiments that were included in this review. The first EBI experiments that taught academically relevant skills to college students were published in 1989, 18 years after Sidman (1971) published his first study discussing equivalences between stimuli. Figure 2 shows that EBI investigations have appeared with increasing frequency in recent years, with nearly half of the body of research appearing after the publication of previous EBI reviews (i.e., Fienup et al., 2011; Rehfeldt, 2011). Although most researchers have discussed instruction in terms of stimulus equivalence (87%; see Table 2), we detected no systematic outcome differences between studies that were couched in terms of stimulus equivalence versus relational frame theory and thus make no further reference to this distinction.

Tables 3 and 4 provide details of the 31 EBI experiments. A total of 680 individuals have participated in EBI research, with 550 of these participants completing EBI tutorials and the remaining participants assigned to non-EBI control conditions. Only a minority of studies reported participant demographic characteristics other than college-student status. The majority of experiments focused on academically relevant learning but were conducted in highly controlled laboratory settings (65%). The remaining protocols were embedded within a formal academic program of study. The reviewed experiments tended to incorporate procedures found to be effective as reported in the basic research literature, such as the one-to-many training structure (one sample throughout all training phases; comparison stimuli change with each phase; Arntzen & Holth, 1997) and the simple-to-complex training protocol (intermixing training relations and derived relation probes; Adams, Fields, &

Fig. 2 Cumulative record displaying the number of college-level EBI articles published between the publication of Sidman's (1971) seminal experiment and August 2016. EBI equivalence-based instruction

Verhave, 1993; Fienup et al., 2015). Most EBI experiments omitted a formal assessment of IOA and treatment integrity; however, the omission of such data could reflect that most studies used automated (computerized) procedures that fully standardize instruction and data recording, thereby making such measures unnecessary.

EBI researchers measured a variety of response topographies and reported a number of different dependent variables (see Tables 3 and 4). Most experiments included MTS procedures in either a computer-based or written format. Researchers included additional response topographies such as writing names (written topographical) of stimuli or vocal naming of stimuli (spoken topographical), which served as measures of response generalization. When reporting the effects of EBI, the vast majority of EBI experiments focused on the effectiveness of EBI in terms of the percentage of correct responses on tests of equivalence class formation (87%) of studies and efficiency as defined by the number of trials (87%) or amount of time (32%) required to form equivalence classes.

Question 1: Is EBI Effective?

The rightmost three columns in Table 1 display the types of comparisons each experiment evaluated, with several experiments evaluating multiple comparisons (e.g., EBI compared to both a no-instruction control condition and an active instruction

Fig. 3 Effect sizes for EBI versus NIC group design experiments. Primary measures are displayed in the top panel with secondary measures in the bottom panel. EBI equivalence-based instruction, MTS match to sample. Albright et al. (2015) (1); Albright et al. (2016) (2); Critchfield (2014) (3); Fienup and Critchfield (2010) (4); Fienup and Critchfield (2011) (5); Fienup et al. (2009), Exp 1 (6); Fienup et al. (2016) (7); Fienup et al. (2015) (8); Lovett et al. (2011) (9); Sandoz and Herbert (2016) (10); Walker and Rehfeldt (2012) (11); Walker et al. (2010), Exp 1 (12) and Exp 2 (13) (top panel). Albright et al. (2015) (1); Albright et al. (2016) (2); Alter and Borrero (2015) (3); Fienup and Critchfield (2011) (4); Fienup et al. (2016) (5); Ninness et al. (2006) (6); Walker and Rehfeldt (2012) (7); Walker et al. (2010), Exp 1 (8); O'Neill et al. (2015) (9); Fields et al. (2009) (10)

condition). We quantified the effects of 23 experiments that asked a basic question about the effectiveness of EBI compared to a baseline or control condition (see Table 1). This category of studies included 18 group design experiments and 5 single-subject design experiments.

Primary Measures Group design analyses included both within- and between-subject measures. For within-subject comparisons, EBI posttest scores were compared to EBI pretest scores. For between-subject comparisons, EBI posttest scores were compared to no-instruction control posttest scores. Figure 3 (top panel) displays Hedges's g values, CIs, and a corresponding forest plot for 13 experiments. Omnibus Hedges's g was 1.59, 95% CI [1.35, 1.82], indicating a large effect of EBI when compared with no instruction for primary measures. Hedges's g values ranged from 0.49, 95% CI [.07, .90], to 8.23, 95% CI [4.94, 11.52]. Effect sizes for 12 individual experiments were large. Only one case had a small effect size—this case was a comparison of pre- and post-computer-based MTS scores in the study by Sandoz and Hebert (2017). This small effect size can be attributed to high pretest scores on the computer-based MTS task (on average only 7% lower than posttest scores), which suggests ceiling effects. An exclusion criterion based on high pretest scores might have increased statistical power by increasing the chance of finding a treatment effect. This outcome contrasts with those of other experiments, which tended to assess pretraining performances at chance-level responding (e.g., 25% with four classes) and equivalence class formation performances between 90% and 100% correct (e.g., Fields et al., 2009).

Figure 4 (top panel) displays single-subject design effect sizes, including a forest plot of omnibus IRD values for the primary measures. Individually, all four studies included in this analysis demonstrated a large effect. Omnibus IRD was 0.95, 95% CI [.64, 1.00], also demonstrating a large effect. One of the four studies demonstrated a moderate effect, whereas the remaining three experiments demonstrated a large effect.

Secondary Measures Figure 3 (bottom panel) displays Hedges's g values, CIs, and a corresponding forest plot for the 10 experiments reporting secondary measures (i.e., response topographies that differed from the training topography). Omnibus Hedges's g was 2.95, 95% CI [2.02, 3.88], indicating a large effect of EBI when compared with no

Fig. 4 Effect sizes for EBI versus NIC single-subject design experiments. Primary measures are displayed in the top panel with secondary measures in the bottom panel. IRD improvement rate difference. Fienup et al. (2010) (1); Reyes-Giordano and Fienup (2015) (2); Sella et al. (2014) (3); Trucil et al. (2015) (4) (top panel). Ninness et al. (2009), Exp 2 (1); Trucil et al. (2015) (2) (bottom panel)

Fig. 5 Effect sizes for EBI versus EBI experiments. All were group designs with only primary measures. EBI equivalence-based instruction, STC simple to complex, SIM simultaneous. Fienup et al. (2015), Exp 1 (1) and Exp 2 (2); Fienup et al. (2016) (3)

instruction for secondary measures. Hedges's g values ranged from 0.94, 95% CI [.26, 1.61], to 5.29, 95% CI [2.63, 7.95]. The secondary measures in these experiments included vocal tests (tact and intraverbal responses regarding stimuli), maintenance measures, and paper-based measures. All measures demonstrated a large effect of EBI on equivalence class formation compared with no-instruction control conditions.

Figure 4 (bottom panel) displays single-subject design effect sizes, including a forest plot of omnibus IRD values for the secondary measures. Both experiments (Ninness et al., 2009; Trucil et al., 2015) included in this analysis showed a large effect. Omnibus IRD for secondary measures was 0.79, 95% CI [.74, .79], demonstrating a large effect.

Question 2: Are There Variations of EBI that Produce Better Academic Outcomes?

Three experiments included in this meta-analysis were comparisons of variations of EBI with a between-subject manipulation (see Table 1). Figure 5 displays Hedges's g values, CIs, and a corresponding forest plot for three experiments. Omnibus Hedges's g was 0.44, 95% CI [.03, .85], indicating a small effect of EBI when compared with another EBI procedure. Fienup et al.'s (2015) comparison of the simple-to-complex and simultaneous training protocols produced a relatively smaller effect with three-member classes (0.40, 95% CI [-0.19, .99]) and a larger effect with four-member classes (0.56, 95% CI [-0.14, 1.27]). Fienup et al.'s (2016) comparisons of training sequences (i.e., which stimulus was the node) across four different classes all had a small effect size of 0.22, 95% CI [-0.60, 1.03].

Outcome	Hedge's g	Lower limit	Upper limit
1 Combined	0.09	-0.63	0.81
2 EBI vs. Reading	0.67	-0.09	1.44
3 Grand Total Omnibus	0.36	-0.16	0.89

Outcome	Hedge's g	Lower limit	Upper limit
1 SE vs CI	0.36	-0.36	1.09
2 Written MTS EBI vs Lecture	0.16	-0.65	0.97
3 Combined	0.44	-0.40	1.28
4 Grand Total Omnibus	0.32	-0.13	0.78

Fig. 6 Effect sizes for EBI versus active control experiments, which were all group designs. Primary measures are displayed in the top panel with secondary measures in the bottom panel. EBI equivalence-based instruction, SE stimulus equivalence, CI complete instruction, MTS match to sample. Fienup and Critchfield (2011) (1); O'Neill et al. (2015) (2) (top panel). Fienup and Critchfield (2011) (1); Lovett et al. (2011) (2); O'Neill et al. (2015) (3) (bottom panel)

Question 3: Is EBI more Effective than Alternative Instructional Strategies?

Primary Measures Three experiments included in this meta-analysis compared EBI with an active control condition using a between-subject manipulation. Figure 6 (top panel) displays Hedges's g values, CIs, and a corresponding forest plot for two experiments. Across two lessons, Fienup and Critchfield (2011) compared EBI outcomes to those following complete instruction (i.e., directly teaching all relations). O'Neill et al. (2015) compared EBI to reading a textbook. Omnibus Hedges's g was 0.36, 95% CI $[-.16, .89]$, indicating a small effect size when comparing EBI to instructional control procedures on primary measures. The omnibus effect size includes a small effect size (Fienup & Critchfield, 2011) and a medium effect size (O'Neill et al., 2015). The small effect size for Fienup and Critchfield (2011) suggests similar levels of student mastery for EBI and a "teach all relations" approach, although EBI was more efficient than the teach all relations approach (i.e., required significantly fewer trials and less training time). Comparisons of selection-based intraverbal responding between an equivalence group and a reading group (O'Neill et al., 2015) showed that EBI has a medium effect size when compared to reading a text. Overall, with so few relevant experiments available, it seems premature to draw any firm conclusions about how the effects of EBI compare with those of other instructional strategies.

Secondary Measures Figure 6 (bottom panel) displays Hedges's g values, CIs, and a corresponding forest plot for two measures across three experiments. This analysis compared EBI to a teach all relations approach (Fienup & Critchfield, 2011), a videotaped lecture (Lovett et al., 2011), and reading a textbook (O'Neill et al., 2015). Omnibus Hedges's g was 0.32, 95% CI $[-.13, .78]$, indicating a small effect of EBI compared with an instructional control procedure on secondary measures. The three experiments included in this analysis each had small effect sizes, and the educational significance of these effects is tentative. EBI participants, on average, required more time to finish instruction than did those who watched a videotaped lecture (Lovett et al., 2011) or read a textbook passage (O'Neill et al., 2015), whereas Fienup and Critchfield (2011) showed that EBI was more efficient than the teach all relations approach.

Discussion

In the past decade, there has been a dramatic increase in the number of published articles that use basic principles of stimulus equivalence in the design of college-level instruction. Effect size calculations for both group and single-subject designs show that EBI is an effective procedure for teaching a wide range of academically relevant concepts to college students. EBI effectively increased class-consistent responding when compared with a preassessment or when compared with a no-instruction control group, and this effect was large and therefore presumably educationally significant. Fewer studies have compared variations of EBI to each other, and to date, no dramatic differences in outcomes have been reported. The same is true for effectiveness comparisons of EBI to active control instruction, although it appears that under at least some circumstances, EBI is more efficient at producing new repertoires. The latter is especially important because efficiency has been the primary basis on which EBI is

recommended (e.g., Critchfield & Fienup, 2008; Critchfield & Twyman, 2014). Additionally, like all behavioral systems of instruction, EBI offers the potential benefit of self-paced, mastery-based, student-driven learning.

Effect sizes calculated in the current meta-analysis should be viewed as preliminary because, as with nearly all reviews, this one does not encompass all possible data sets. Several relevant articles have been published since the closing of our data collection window (e.g., Fienup & Brodsky, 2017; Greville, Dymond, & Newton, 2016; Varelas & Fields, 2017), and we could not obtain raw data for some investigations that were in print when the analysis was conducted and additionally, all reviews confront a potential file-drawer problem involving the omission of unpublished studies; note that for the present report, we chose to focus only on published studies that had been evaluated for quality in peer review.

Inclusion of additional experiments may well have changed our conclusions, particularly in analyses that only included a few studies.

For example, our meta-analysis did not include Zinn, Newland, and Ritchie (2015), one of the most promising EBI experiments, because we were unable to obtain raw data. Zinn et al. (2015) compared an EBI program to a criterion-control group, in which participants practiced relations drawn at random from the EBI stimulus set, and a trial-control group, in which the number of trials participants completed was yoked to the number of trials EBI participants completed. Zinn et al. (2015) found superior effects for EBI. Inclusion of this study would enhance support for the effectiveness of EBI compared with other instructional interventions—support that in the present review was based on limited evidence and appeared to be modest in magnitude. If our review serves no other purpose, it may be to highlight that EBI research remains in an emerging phase, and additional experiments comparing EBI to active instructional controls are desperately needed if this technology is to be adopted outside of behavior science and in college classrooms.

Validity Issues

The results of the current systematic review and meta-analysis can help guide future applied, college-level EBI experiments by focusing on three concepts that affect experimental decisions: internal validity, statistical conclusion validity, and external validity.

Internal Validity The EBI evidence base consists of experiments that use a variety of research designs. Most experiments implemented group designs, and a growing number of experiments have evaluated the effects of EBI using multiple-baseline, single-subject experimental designs.³ Various research designs control for threats to internal validity in different ways. For example, the multiple-baseline design controls for threats such as history, maturation, and testing by repeatedly measuring behavior before and after EBI and staggering the onset of EBI across participants or classes (Baer, Wolf, & Risley, 1968). A number of the group design experiments identified by our search would be categorized as quasi-experimental by Campbell and Stanley (1963). For example, Fienup and Critchfield (2010) and Albright et al. (2016) exposed 10 and 11

³ The time series design as discussed by Campbell and Stanley (1963) does not reflect the experimental rigor of modern single-subject designs—which were developed after the publication of their book—that include reversals and staggered baselines to control for threats to internal validity. Thus, although Campbell and Stanley categorize time series designs as quasi-experimental, we contend that single-subject designs identified by this search represent well-controlled experimental designs.

participants, respectively, to an instructional sequence that included pretesting of classes, EBI, and post-EBI class formation tests. This type of design is most useful during the initial development of EBI for a particular content area. The researchers who conduct these experiments do so in laboratory settings with participants who have no prior experience with the content, and an individual's experience is completed in one session, thus controlling for the influence of outside instruction. However, pre-post designs do not control for a number of other threats to internal validity, such as testing and instrumentation (Campbell & Stanley, 1963).

Only a subset of the group designs we examined can be categorized as “true experimental” designs according to Campbell and Stanley (1963; see Table 3). One such experiment, conducted by Lovett et al. (2011), compared EBI with a videotaped lecture condition and reduced most threats to internal validity by making preintervention and postintervention assessments of equivalence class formation for participants completing EBI and lecture instructional formats. Fields et al. (2009) also included preintervention assessments in between-group comparisons and added a no-instruction control group to demonstrate that the passage of time or exposure to tests alone does not improve class-consistent responding.

As EBI moves out of the laboratory setting, it is imperative that researchers control for threats to internal validity that become increasingly problematic in naturalistic instructional environments, such as the threat of outside educational influence. Choosing group or single-subject experimental designs that control for such internal validity threats in naturalistic educational settings should be an important procedural decision in future EBI research.

Statistical Conclusion Validity Statistical conclusion validity is the degree to which statistically significant mean differences can be detected where they exist (Cook & Campbell, 1979). Increasing and demonstrating high statistical conclusion validity in future research can help increase the adoption of EBI by educators who do not specialize in behavior science. Although there are issues with hypothesis testing (e.g., type I and II errors, low power), other fields of psychology rely on these types of data to make treatment decisions. Furthermore, although behavior scientists are well acquainted with baseline logic procedures, this practice may not be commonly accepted by others. Providing effect sizes for EBI treatments can help demonstrate the magnitude of treatment both within and outside the field.

Many experiments in this analysis increased statistical conclusion validity in three important ways. First, researchers collected multiple performance measures before, during, and after equivalence class formation, often in addition to between-subject comparisons, to demonstrate powerful treatment effects while conserving resources, such as extensive use of a participant subject pool and experimenter time, allowing researchers to conduct further experiments with the saved resources. Second, researchers maximized statistical conclusion validity by making treatment differences as large as possible for EBI versus control participants by omitting instruction for control participants (e.g., Fields et al., 2009). Third, researchers eliminated participants who scored high on preintervention measures to prevent ceiling effects and to ensure that pretest and posttest measures were as different as possible (e.g., Ninness et al., 2006).

Across the experiments included in this analysis, data displays showed individual data points and averaged scores with corresponding variability measures. These data displays allow readers to see the effects of EBI at both the group and individual levels. A number of experiments included in this meta-analysis provided inferential statistics to support

conclusions from visual analyses. For example, Sandoz and Hebert (2017) reported the results of a paired-samples *t* test to show that posttest outcomes were different from pretest outcomes due to treatment rather than sampling error. O'Neill et al. (2015) reported the results of a multivariate analysis of covariance (MANCOVA) that examined whether there were statistically significant differences between an EBI group and a reading group on a variety of dependent measures. The statistical outcomes reported by researchers verified the differences that were apparent through visual analysis. Only a few studies reported effect size calculations (e.g., Fields et al., 2009; Fienup & Critchfield, 2011). Reporting effect sizes may encourage instructors and practitioners with backgrounds in traditional experimental methodology to adopt EBI pedagogy. Publishing inferential statistics along with data displays emphasizing individual data may promote wider acceptance of EBI, improving the social validity of behavior-analytic procedures for other subfields of psychology and education and helping disseminate this effective technology.

External Validity and Comparative Effectiveness EBI experiments in this analysis demonstrated external validity in important ways. Across all experiments, 550 individuals, and several different content areas, EBI demonstrated generally strong effects. Fienup et al. (2016) showed that completing 1 h of EBI of outside class time produced a gain of 23 percentage points, on average, on a classroom examination (Hedges's $g = 3.69$, 95% CI [2.57, 4.82]). EBI helped participants increase class-consistent responding across response topographies commonly assessed in higher education such as multiple-choice tests (Albright et al., 2016; Walker et al., 2010) but also produced increases in advanced repertoires such as talking and writing about the educational stimuli (Lovett et al., 2011; Walker et al., 2010). Pytte and Fienup (2012) demonstrated that lecturing using EBI also leads to instructional gains in a naturalistic lecture setting. Walker and Rehfeldt (2012), Critchfield (2014), and O'Neill et al. (2015) demonstrated EBI's efficacy through course management systems, which allow students to engage in course material outside class time, and Walker et al. (2011) demonstrated that a worksheet format could be used to deliver EBI. Collectively, the data presented in this analysis demonstrate the generality of EBI across many different individuals, a number of content areas, topographies of responding, different contexts of implementation, and different formats of implementation.

The number of participants who have benefited from EBI is impressive, yet little is known about specific participant characteristics. Researchers should make efforts to expand EBI's generality across populations. A first step to accomplishing this goal is to provide a thorough report on participant demographics, including both academic and cultural information (e.g., Sella et al., 2014). Researchers can also expand participant diversity by conducting research with a wide variety of students, perhaps by collaborating with other researchers working at culturally and economically diverse institutions. Other ways to accomplish this objective include using students at various education levels, such as graduate students (e.g., Walker & Rehfeldt, 2012), and expanding the age range of participants beyond the typical 18- to 23-year-old demographic (e.g., Albright et al., 2016).

While existing evidence supports the effectiveness of EBI, Rehfeldt (2011) suggested that many of the relevant studies qualify as demonstration-of-concept exercises. To date, the effects of EBI have rarely been studied in a context representing the amount and breadth of content a student is required to learn in a college course. Recent experiments (e.g., Greville et al., 2016; not included in the meta-analysis because this study was published after the systematic review article identification process) have scaled up

instruction to teach more (and larger) stimulus classes. This direction is promising, as most EBI experiments are carried out with a small number of stimuli (e.g., four three- or four-member classes) that represent only a fraction of what is taught in a semester-long college course. The effectiveness of Greville et al.'s (2016) study, combined with learning set effects (Fienup et al., 2016), indicate that EBI has the potential to show large instructional gains if it is used throughout an entire course. Furthermore, it is unknown whether best practices for EBI, as established based on basic and translational research, translate to the applied context in which students learn course-relevant content. It is important to determine whether procedural variations identified as most effective in basic research contexts are still most effective in applied contexts to test whether context variables in applied settings obfuscate outcome differences between procedural variations. If differences between procedural variations are not apparent in applied contexts, then when instructors are designing EBI for classroom use, they should program for the procedural variation that requires less response effort on the part of the instructor. For example, Narthey, Arntzen, and Fields (2015) determined that the sequence of training stimuli affects equivalence class formation in the basic context, but Fienup et al. (2016) were not able to replicate these effects in the applied context. Arntzen (2004) identified the linear series training structure as least effective, but applied studies such as those by Fields et al. (2009) and Fienup et al. (2016) used this structure with success. With such instructional variables, their effects may be lessened or not present when investigated in naturalistic educational settings and the context variables present there.

A number of experiments represented in this meta-analysis taught content that is directly relevant to psychology students, such as statistics (e.g., Albright et al., 2016), and researchers have expanded to novel non-psychology content areas, such as mathematics (e.g., Ninness et al., 2006) and portion size estimation (Hausman et al., 2014). EBI research in content areas outside psychology could help students learn material for college classes that are notoriously difficult, such as organic chemistry, physics, and calculus (for application to the teaching of neuroanatomy, see Fienup et al. (2010), Fienup et al. (2016), and Pytte and Fienup (2012)). Experiments demonstrating EBI's effectiveness in a naturalistic setting may increase its generality and help push this technology toward mainstream use.

More research is needed to compare EBI to traditional instructional methods and some of the experiments that have made such comparisons warrant clarification. For example, Lovett et al. (2011) found that EBI conducted in a quiet laboratory setting was more effective than watching a video lecture in the same quiet laboratory setting, but this effect had a small educational significance. Fienup and Critchfield (2011) found that EBI was more effective than learning all relations in stimulus classes, but the educational significance was also small. Although these comparisons are useful toward developing a technology of EBI, these comparison conditions do not necessarily represent traditional instructional methods delivered in naturalistic settings with the accompanying distractions—the experiments implemented controlled, laboratory versions of both EBI and “typical instruction.” In the naturalistic setting, EBI users may collaborate with peers or engage in alternate behaviors (e.g., phone and Internet use) while completing, or failing to engage with, instruction. O'Neill et al. (2015) compared EBI and reading through an online course management system that allowed students to work at preferred times in their desired settings. In other words, O'Neill and colleagues tested EBI in a context similar to what students enrolled in that course experienced and found a positive effect of EBI relative to reading a text. Varelas and Fields (2017; not published at the time of the

systematic review process) ventured into the classroom to determine the effectiveness of clicker technology on the formation of equivalence classes to students enrolled in a lifespan development course. Conducting further studies in the classroom setting will contribute to the evidence base of EBI's efficacy and social validity and will help mainstream its use in college and university settings.

Toward a Highly Applicable Research Agenda in Higher Education

Much of the EBI research has been conducted in highly controlled settings and thus may best be conceptualized as experimental analysis of behavior or translational research (Mace & Critchfield, 2010). The technology shows great promise when used under controlled conditions using research volunteer participants and teaching a few three- or four-member equivalence classes. In the last few years, some researchers have stepped out of the lab to evaluate EBI in more naturalistic settings (Greville et al., 2016; O'Neill et al., 2015; Pytte & Fienup, 2012; Varelas & Fields, 2017). Future research needs to focus more directly on application and a number of research questions that will arise as EBI researchers tackle scaled-up curricula in naturalistic college settings.

First, researchers need to find ways to incorporate EBI into college classrooms. Does EBI replace lecturing (Lovett et al., 2011), change how an instructor lectures (Pytte & Fienup, 2012), or supplement typical classroom activities (Fienup et al., 2016)? Researchers have conducted studies demonstrating EBI's use in each of these ways; however, for an instructor interested in adopting EBI, it may be unclear how to incorporate EBI technology into the classroom because there are no examples of a fully integrated EBI curriculum in the literature. In naturalistic educational settings, questions remain regarding the structuring of EBI across an entire semester and contingencies that promote the continued use of EBI tutorials. One potential application of EBI is to combine EBI with interteaching, a behavior-analytic pedagogy with considerable research support that includes prep guides, contingencies for completing small-group discussions on course material, and supplemental lectures to clarify remaining content questions (Boyce & Hinline, 2002; Sturmey, Dalfen, & Fienup, 2015). EBI could be used to teach basic concepts to mastery before students complete prep guides for classroom interteaching sessions. Ultimately, research that brings EBI out of the laboratory setting to compare EBI to other instructional strategies in their own naturalistic settings will help answer questions regarding whether EBI is better than other methods or instead most useful when combined with other methods. The results of the present meta-analysis suggest that EBI may produce a small benefit compared with other pedagogies, which may not be enough to prompt instructors to change from teaching as usual to EBI given the current response effort of setting up EBI tutorials for a course.

Second, researchers should address the dissemination of this technology. Studies such as those by Walker and Rehfeldt (2012) and Critchfield (2014) have shown that it is possible to use common online learning tools to deliver EBI to students, whereas Walker et al. (2010) administered paper-based worksheets. Classroom instructors could benefit from task analyses for implementing EBI in the classroom using resources that are readily available. Additionally, developing the technology for mobile devices could boost EBI's dissemination. Such applications would allow students to complete instruction on their phones and tablets. Students could complete EBI on the go—while traveling, while

commuting, or while waiting for appointments—thereby maximizing the student's instructional time. Comparative effectiveness experiments conducted in naturalistic educational settings and the development of tools enabling the use of EBI in classrooms may ultimately facilitate the adoption of EBI as a common pedagogy in higher education.

Third, basic stimulus equivalence research has influenced the development of EBI applications. Large-scale, naturalistic evaluations of EBI allow for clarification of basic principles, thus informing basic science. Testing EBI in the naturalistic environment may help identify which variables maximize EBI's effectiveness and which principles are less relevant due to the differences in participants, motivational characteristics, and stimuli. Fienup et al. (2016) tested the effects of stimulus order on equivalence class formation with students enrolled in a behavioral neuroscience course who learned neuroanatomy concepts from that course. The researchers did not find an effect of stimulus order, contradicting findings in basic research settings (Arntzen, 2004; Nartey et al., 2015). In applied settings, research can examine the generality of basic research findings and provide clarification on how additional variables (e.g., motivational characteristics) influence other established functional relations (e.g., effect of stimulus order). Many variables have yet to be evaluated in applied contexts, such as stimulus salience (Arntzen, 2004) and training structure. Across a semester of EBI tutorials, one could begin to examine how learning set (Fienup et al., 2016) attenuates differences between EBI manipulations, such as training protocol. This examination would help to clarify the variables that affect the formation of educationally relevant equivalence classes with students who experience all of the competing contingencies that regularly occur in higher education settings and are (potentially) motivated to learn the content.

Collectively, findings from the current systematic review and meta-analysis show that EBI technology is effective and warrants further investigation. The effectiveness of EBI and learners' positive evaluations of its use (Fienup & Critchfield, 2011; Greville et al., 2016) demonstrate the need for optimizing, expanding, and disseminating this technology to facilitate academic success based on positive reinforcement rather than aversive control. However, the long-term fate of EBI rests on research that has yet to be conducted. To date, although the data have not always shown that EBI is more effective than mainstream instructional methods such as lecture, studies have shown its efficiency relative to treatment as usual. The relative efficiency of EBI can help both students and instructors at the college level, but instructors may require demonstrations of large differences in effectiveness and/or efficiency in the classroom setting before they are willing to adopt this technology. We can reduce the cost of adopting this technology by producing task analyses or video models for creating EBI tutorials using the resources available to college instructors (e.g., Blackboard access). Of utmost importance are empirical demonstrations that EBI produces better educational outcomes in less time than the pedagogy instructors currently use. These data could ultimately lead to the adoption of EBI as a standard pedagogy in higher education.

Author Note The first author completed this study in partial fulfillment of a doctoral degree in Psychology through the Graduate Center, CUNY. A portion of Dr. Fienup's work was completed while he was affiliated with Queens College, CUNY.

Acknowledgments We thank Ria Bissoon, Haeri Gim, Radiyyah Hussein, and Rika Ortega for assistance in conducting this study. We thank Drs. Alexandra Logue and Robert Lanson for comments on an earlier version

of this manuscript. We also thank the following researchers for providing raw data sets for this study: Dr. Leif Albright, Dr. Thomas Critchfield, Dr. John O'Neill, Dr. Kenneth Reeve, Dr. Ruth Anne Rehfeldt, Dr. Emily Sandoz, and Brooke Walker.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

References

References marked with an asterisk indicate studies included in the meta-analysis.

- Adams, B. J., Fields, L., & Verhave, T. (1993). Effects of test order on intersubject variability during equivalence class formation. *The Psychological Record*, 43, 133–152.
- *Albright, L., Reeve, K. F., Reeve, S. A., & Kisamore, A. N. (2015). Teaching statistical variability with equivalence-based instruction. *Journal of Applied Behavior Analysis*, 48, 883–894.
- *Albright, L., Schnell, L., Reeve, K. F., & Sidener, T. M. (2016). Using stimulus equivalence-based instruction to teach graduate students in applied behavior analysis to interpret operant functions of behavior. *Journal of Behavioral Education*, 25, 290–309.
- *Alter, M. M., & Borrero, J. C. (2015). Teaching generatively: learning about disorders and disabilities. *Journal of Applied Behavior Analysis*, 48, 376–389.
- Arntzen, E. (2004). Probability of equivalence formation: familiar stimuli and training sequence. *The Psychological Record*, 54, 275–291.
- Arntzen, E., & Holth, P. (1997). Probability of stimulus equivalence as a function of training design. *The Psychological Record*, 47, 309–320.
- Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis*, 1, 91–97.
- Boyce, T. E., & Hineline, P. N. (2002). Interteaching: a strategy for enhancing the user-friendliness of behavioral arrangements in the college classroom. *The Behavior Analyst*, 25, 215–226.
- Bureau of Labor Statistics. (2016). *Labor market activity, education, and partner status among America's young adults at 29: results from a longitudinal survey*. Retrieved from <https://www.bls.gov/news.release/pdf/nlsyth.pdf>
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research: handbook of research on teaching*. Chicago, IL: Rand McNally.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 12, 997–1003.
- Comprehensive Meta-Analysis (Version 3.3) [Computer software]. (2014). Englewood, NJ: Biostat Available from <http://www.meta-analysis.com>.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: design and analysis for field settings*. Boston, MA: Houghton Mifflin.
- *Critchfield, T. S. (2014). Online equivalence-based instruction about statistical inference using written explanations instead of match-to-sample training. *Journal of Applied Behavior Analysis* 47, 606–611.
- Critchfield, T. S., & Fienup, D. M. (2008). Stimulus equivalence. In S. F. Davis & W. F. Buskist (Eds.), *21st century psychology* (pp. 360–372). Thousand Oaks, CA: Sage.
- *Critchfield, T. S., & Fienup, D. M. (2010). Using stimulus equivalence technology to teach statistical inference in a group setting. *Journal of Applied Behavior Analysis*, 43, 763–768.
- *Critchfield, T. S., & Fienup, D. M. (2013). A “happy hour” effect in translational stimulus relations research. *The Experimental Analysis of Human Behavior Bulletin*, 29, 2–7.
- Critchfield, T. S., & Twyman, J. S. (2014). Prospective instructional design: establishing conditions for emergent learning. *Journal of Cognitive Education and Psychology*, 13, 201–217.
- Ellis, P. D. (2010). *The essential guide to effect sizes: statistical power, meta-analysis, and the interpretation of research results*. Cambridge: Cambridge University Press.
- Field, A. (2009). *Discovering statistics using SPSS*. Thousand Oaks, CA: Sage.

- *Fields, L., Travis, R., Roy, D., Yadlovker, E., de Aguiar-Rocha, L., & Sturmey, P. (2009). Equivalence class formation: a method for teaching statistical interactions. *Journal of Applied Behavior Analysis*, 42, 575–593.
- Fields, L., Verhave, T., & Fath, S. (1984). Stimulus equivalence and transitive associations: a methodological analysis. *Journal of the Experimental Analysis of Behavior*, 42, 143–157.
- Fienup, D. M., & Brodsky, J. (2017). Effects of mastery criterion on the emergence of derived equivalence relations. *Journal of Applied Behavior Analysis*, 50, 843–848.
- *Fienup, D. M., Covey, D. P., & Critchfield, T. S. (2010). Teaching brain–behavior relations economically with stimulus equivalence technology. *Journal of Applied Behavior Analysis*, 43, 19–33.
- *Fienup, D. M., & Critchfield, T. S. (2010). Efficiently establishing concepts of inferential statistics and hypothesis decision making through contextually controlled equivalence classes. *Journal of Applied Behavior Analysis*, 43, 437–462.
- *Fienup, D. M., & Critchfield, T. S. (2011). Transportability of equivalence-based programmed instruction: efficacy and efficiency in a college classroom. *Journal of Applied Behavior Analysis*, 44, 435–450.
- *Fienup, D. M., Critchfield, T. S., & Covey, D. P. (2009). Building contextually-controlled equivalence classes to teach about inferential statistics: a preliminary demonstration. *Experimental Analysis of Human Behavior Bulletin*, 27, 1–10.
- Fienup, D. M., Hamelin, J., Reyes-Giordano, K., & Falcomata, T. S. (2011). College-level instruction: derived relations and programmed instruction. *Journal of Applied Behavior Analysis*, 44, 413–416.
- *Fienup, D. M., Mylan, S. E., Brodsky, J., & Pytte, C. (2016). From the laboratory to the classroom: the effects of equivalence-based instruction on neuroanatomy competencies. *Journal of Behavioral Education*, 25, 143–165.
- *Fienup, D. M., Wright, N. A., & Fields, L. (2015). Optimizing equivalence-based instruction: effects of training protocols on equivalence class formation. *Journal of Applied Behavior Analysis*, 48, 1–19.
- Green, G., & Saunders, R. R. (1998). Stimulus equivalence. In K. A. Lattal & M. Perone (Eds.), *Handbook of research methods in human operant behavior* (pp. 229–262). New York, NY: Plenum Press.
- Greville, W. J., Dymond, S., & Newton, P. M. (2016). The student experience of applied equivalence-based instruction for neuroanatomy teaching. *Journal of Educational Evaluation for Health Professions*, 13, 32.
- *Hausman, N. L., Borrero, J. C., Fisher, A., & Kahng, S. (2014). Improving accuracy of portion-size estimations through a stimulus equivalence paradigm. *Journal of Applied Behavior Analysis*, 47, 485–499.
- *Hayes, L. J., Thompson, S., & Hayes, S. C. (1989). Stimulus equivalence and rule following. *Journal of the Experimental Analysis of Behavior*, 52, 275–291.
- Hayes, S. C., Barnes-Holmes, D., & Roche, B. (2001). *Relational frame theory: a post-Skinnerian account of human language and cognition*. New York, NY: Plenum Press.
- Keller, F. S. (1968). Good-bye, teacher. *Journal of Applied Behavior Analysis*, 1, 79–89.
- Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). Single-case design technical documentation. Retrieved from <https://ies.ed.gov/ncee/wwc/Document/229>
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.
- *Lovett, S., Rehfeldt, R. A., Garcia, Y., & Dunning, J. (2011). Comparison of a stimulus equivalence protocol and traditional lecture for teaching single-subject designs. *Journal of Applied Behavior Analysis*, 44, 819–833.
- Mace, F. C., & Critchfield, T. S. (2010). Translational research in behavior analysis: historical traditions and imperative for the future. *Journal of the Experimental Analysis of Behavior*, 93, 293–312.
- *McGinty, J., Ninness, C., McCuller, G., Rumph, R., Goodwin, A., Kelso, G., ... Kelly, E. (2012). Training and deriving precalculus relations: a small-group, web-interactive approach. *The Psychological Record*, 62, 225–242.
- Michael, J. (1991). A behavioral perspective on college teaching. *The Behavior Analyst*, 14, 229–239.
- Mulryan-Kyne, C. (2010). Teaching large classes at college and university level: challenges and opportunities. *Teaching in Higher Education*, 15, 175–185.
- Nartey, R. K., Arntzen, E., & Fields, L. (2015). Training order and structural location of meaningful stimuli: effects of equivalence class formation. *Learning & Behavior*, 43, 342–353.
- *Ninness, C., Barnes-Holmes, D., Rumph, R., McCuller, G., Ford, A. M., Payne, R. ... Elliott, M. P. (2006). Transformations of mathematical and stimulus functions. *Journal of Applied Behavior Analysis*, 39, 299–321.
- *Ninness, C., Dixon, M., Barnes-Holmes, D., Rehfeldt, R. A., Rumph, R., McCuller, G. ... McGinty, J. (2009). Constructing and deriving reciprocal trigonometric relations: a functional analytic approach. *Journal of Applied Behavior Analysis*, 42, 191–208.

- *O'Neill, J., Rehfeldt, R. A., Ninness, C., Munoz, B. E., & Mellor, J. (2015). Learning Skinner's verbal operants: comparing an online stimulus equivalence procedure to an assigned reading. *The Analysis of Verbal Behavior*, 31, 255–266.
- Parker, R. I., Vannest, K. J., & Brown, L. (2009). The improvement rate difference for single-case research. *Exceptional Children*, 75, 135–150.
- *Pytte, C. L., & Fienup, D. M. (2012). Using equivalence-based instruction to increase efficiency in teaching neuroanatomy. *The Journal of Undergraduate Neuroscience Education*, 10, A125–A131.
- Rehfeldt, R. A. (2011). Toward a technology of derived stimulus relations: an analysis of articles published in the *Journal of Applied Behavior Analysis*, 1992–2009. *Journal of Applied Behavior Analysis*, 44, 109–119.
- *Reyes-Giordano, K., & Fienup, D. M. (2015). Emergence of topographical responding following equivalence-based neuroanatomy instruction. *The Psychological Record*, 65, 495–507.
- *Sandoz, E. K., & Hebert, E. R. (2017). Using derived relational responding to model statistics learning across participants with varying degrees of statistics anxiety. *European Journal of Behavior Analysis*, 18, 113–131.
- *Sella, A. C., Ribeiro, D. M., & White, G. W. (2014). Effects of an online stimulus equivalence teaching procedure on research design open-ended questions performance of international undergraduate students. *The Psychological Record*, 64, 89–103.
- Sidman, M. (1971). Reading and auditory-visual equivalences. *Journal of Speech, Language, and Hearing Research*, 14, 5–13.
- Sidman, M. (1994). *Equivalence relations and behavior: a research story*. Boston, MA: Authors Cooperative.
- Sidman, M., & Tailby, W. (1982). Conditional discrimination vs. matching to sample: an expansion of the testing paradigm. *Journal of the Experimental Analysis of Behavior*, 37, 5–22.
- Skinner, B. F. (1968). *The technology of teaching*. East Norwalk, CT: Appleton-Century-Crofts.
- Sturmey, P., Dalfen, S., & Fienup, D. M. (2015). Inter-teaching: a systematic review. *European Journal of Behavior Analysis*, 16, 121–130.
- *Trucil, L. M., Vladescu, J. C., Reeve, K. F., DeBar, R. M., & Schnell, L. K. (2015). Improving portion-size estimation using equivalence-based instruction. *The Psychological Record*, 65, 761–770.
- Varelas, A., & Fields, L. (2017). Equivalence based instruction by group based clicker training and sorting tests. *The Psychological Record*, 67, 71–80.
- *Walker, B. D., & Rehfeldt, R. A. (2012). An evaluation of the stimulus equivalence paradigm to teach single-subject design to distance education students via blackboard. *Journal of Applied Behavior Analysis*, 45, 329–344.
- *Walker, B. D., Rehfeldt, R. A., & Ninness, C. (2010). Using the stimulus equivalence paradigm to teach course material in an undergraduate rehabilitation course. *Journal of Applied Behavior Analysis*, 4, 615–633.
- *Zinn, T. E., Newland, M. C., & Ritchie, K. E. (2015). The efficiency and efficacy of equivalence-based learning: a randomized controlled trial. *Journal of Applied Behavior Analysis*, 48, 865–882.